DOCUMENT RESUME

ED 458 821 FL 026 977

AUTHOR Nakamura, Yuji
TITLE Rasch Based Analysis of Oral Presentation Assessment for Item Banking.
PUB DATE 2001-09-00
NOTE 13p.; This research was supported in part by Tokyo Keizai University under research grant CPU02-00.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Item Response Theory; Language Tests; *Oral Language; *Second Language Instruction; Second Language Learning; *Student Evaluation; Test Items; *Test Validity; Testing
IDENTIFIERS Rasch Model

ABSTRACT

The Rasch Model is an item response theory, one parameter model developed that states that the probability of a correct response is a function of the difficulty of the item and the ability of the candidate. Item banking is useful for language testing. The Rasch Model provides estimates of item difficulties that are meaningful, irrespective of ability level tested. Application of a many-faceted Rasch measurement model is one of the current issues in the language testing area. This paper explores what the Rasch model tells us in analyzing the multifaceted data of assessment in presentation classes, and focuses more specifically on the possible ways of constructing item banks for oral presentation classes so that teachers will know how difficult a set of test items is for the student-raters and how well those items can distinguish between the better and the poor students. This paper explores several areas of importance with respect to assessment, specifically, what the many-faceted Rasch Measurement Model tells us about the following factors: the relationship among the three facets of assessment (students, items, raters); rater severity/leniency; and students' ability. All of these questions are answered with extensive empirical basis for judgment. (KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 458 821

Rasch Based Analysis of Oral Presentation Assessment for Item Banking

Yuji Nakamura

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Yuji Nakamura

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

人文自然科学論集　第112号　所載抜刷
(2001年9月)

BEST COPY AVAILABLE

論 文

Rasch Based Analysis of Oral Presentation Assessment for Item Banking

Yuji Nakamura

Abstract

Application of a many-facet Rasch measurement (FACETS) model is one of the current issues in the language testing area. This paper explores 1) what the Rasch model tells us in analyzing the multi-faceted data of assessment in presentation classes, and 2) focuses more specifically on possible ways of constructing item banks for oral presentation classes.

1. Purpose of the research

The Rasch measurement model is a powerful tool for handling polytomous data involving raters' judgments (Linacre 1989, 1994). The present paper first explores what the Rasch model tells us in analyzing the multi-faceted data of assessment in presentation classes.

Second, this paper focuses more specifically on possible ways of constructing item banks for oral presentation classes so that teachers will know how difficult a set of test items is for the student-raters and how well those items can distinguish between the better and the poor students.

2. Research questions

This paper explores several areas of importance with respect to assessment:

1) what the Many-Facet Rasch Measurement Model tells us about the following factors:

a. the relationship among the three facets of assessment (students, items, raters)

b. rater severity/leniency

c. students' ability

d. item difficulty (and/or discrimination)

e. the function of rating categories

2) how the assessment or the test can be improved by utilizing the analyzed data

3) how item banks can be constructed using the Rasch model

3. Research design and methods

In class, twelve students gave public speaking presentations that were assessed by five raters (one classroom teacher and four students, who were chosen at random) using eleven evaluation items (e. g., sincerity, eye contact, and oral fluency). Three facets of the study--student ability, item difficulty and rater severity--plus rating categories, will be thoroughly discussed.

1) Subjects:

Twelve university students

2) Raters:

Five raters (one teacher and four students who were chosen at random from among the twelve students above)

3) Rating items:

A relevant selection of Tatum's items (1997) was chosen and arranged for the present research.

1. speakers' sincerity toward the audience
2. oral fluency
3. pronunciation (clarity or enunciation)
4. eye contact
5. facial expression
6. appropriate language (grammar)
7. originality of expression
8. content (target of the speech)
9. written fluency (smooth flow of ideas)
10. appropriate evidence
11. holistic evaluation (overall impression)

4) Rating scale

Items 1 through 10 in the rating list were scored on a six-point scale (1 was poor and 6 was good), while only item 11 was judged on a four-point scale (1 was poor and 4 was good).

4. Data analysis

The data were analyzed using the Many-Facet Rasch Measurement Model, which was able to give detailed information about three aspects of the study (student ability, item difficulty, and rater severity). The data were investigated mainly from the viewpoint of unexpected scores and fit statistics. Also, the benchmark for the acceptable range of the infit and outfit statistic was set between 0.6-1.4 since this was performance speech test data that involved raters' judgments. Furthermore, the Separation index for the students' measurement report should be over 2.0 in theory.

5. Results and discussion

5.1. What does the Many-Facet Rasch Measurement model tell us about the test facets (students, items, raters) and rating categories?

First let us look at the unexpected responses in Table 1.

Table 1 Unexpected Responses

Cat	Step	Exp.	Resd	StRes	N	r	Nu	st	Nu	items
2	2	3.7	−1.7	−3	1	1	12	12	11	holistic evaluation
6	6	3.7	2.3	3	2	2	5	5	9	written fluency
6	6	3.8	2.2	3	2	2	6	6	5	facial expression
Cat	Step	Exp.	Resd	StRes	N	r	Nu	st	Nu	items

Table 1 shows three unexpected responses. In the first case, rater 1, student 12 and item 11 are related to the result in score 2, whose expected value is 3.7. In the second case, rater 2, student 5 and item 9 interrelate to produce score 6, whose expected value is 3.7. Furthermore, in the third case, rater 2, student 6 and item 5 interact to produce a score of 6, but the expected value is 3.8. Although it is not easy to determine the cause of the discrepancy between the observed scores and the expected scores in these cases, rater 2 may have something to do with this phenomenon because of his frequent appearance in this unexpected data. Thus, this table of unexpected responses can lead us to a further investigation of significant data.

Now let us examine the raters' measurement in Table 2.

Table 2 shows the raters' measurement report. According to our benchmark for the fit statistic of the acceptable range (0.6-1.4) for this research, where raters' judgment is involved in a speaking performance test, all the raters are working rather reasonably--except rater 2, whose infit statistic is 1.5, a value beyond the maximum range (1.4). When we look at the measure column, rater 1 (the teacher) is the most lenient, followed by rater 4, while rater 2 is the severest among the five raters.

Table 2 raters Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S. E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N raters
611	132	4.6	4.65	1.05	.11	1.3	2	1.3	2	1 1
459	132	3.5	3.47	−.93	.13	1.5	3	1.4	2	2 2
498	132	3.8	3.75	−.36	.12	.6	−3	.6	−3	3 3
576	132	4.4	4.36	.64	.11	.8	−1	.9	−1	4 4
495	132	3.8	3.73	−.40	.12	.7	−2	.7	−2	5 5
527.8	132.0	4.0	3.99	.00	.12	1.0	− .4	1.0	− .4	Mean (Count : 5)
56.5	.0	.4	.44	.73	.01	.3	2.7	.3	2.7	S. D.

RMSE (Model) .12 Adj S. D. .72 Separation 6.18 Reliability .97
Fixed (all same) chi-square : 200.0 d. f. : 4 significance : .00
Random (normal) chi-square : 4.0 d. f. : 3 significance : .26

Table 3 students Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S. E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	students
221	55	4.0	3.98	.57	.18	1.3	1	1.4	1	1	1
205	55	3.7	3.69	.05	.19	.7	−1	.7	−1	2	2
217	55	3.9	3.91	.44	.18	.9	0	.9	0	3	3
196	55	3.6	3.54	−.27	.19	.8	−1	.8	−1	4	4
227	55	4.1	4.10	.75	.18	1.1	0	1.1	0	5	5
243	55	4.4	4.41	1.23	.17	1.6	2	1.7	3	6	6
218	55	4.0	3.93	.47	.18	1.0	0	.9	0	7	7
246	55	4.5	4.48	1.32	.17	.6	−2	.6	−2	8	8
227	55	4.1	4.10	.75	.18	.8	−1	.8	−1	9	9
197	55	3.6	3.55	−.23	.19	1.2	0	1.1	0	10	10
206	55	3.7	3.71	.08	.18	.8	0	.8	−1	11	11
236	55	4.3	4.27	1.02	.17	.9	0	1.0	0	12	12
219.9	55.0	4.0	3.97	.51	.18	1.0	− .2	1.0	− .2	Mean (Count : 12)	
16.1	.0	.3	.30	.51	.01	.3	1.4	.3	1.5	S. D.	

RMSE (Model) .18 Adj S. D. .48 Separation 2.66 Reliability .88
Fixed (all same) chi-square : 95.5 d. f. : 11 significance : .00
Random (normal) chi-square : 11.0 d. f. : 10 significance : .36

It should also be pointed out that the Separation index, 6.18, is a bit high, which means that the raters' judgments vary greatly. However, the student raters as a whole do good jobs, with all at a rather consistent level of severity. This is shown by the fit statistic.

The cause of the misfit of rater 2 should also be examined. It might be very difficult to relate his severity to the misfit result ; however, it is worth trying to find a reasonable explanation, because rater 2 is highly involved in two of three unexpected responses in Table 1 shown above.

Now let us go on to the student measurements in Table 3.

Table 3 presents the student measurement report. In other words, it shows student ability. The measure column indicates that student 8 is the most able, followed by student

Table 4 items Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu items
256	60	4.3	4.22	.22	.17	.9	0	.8	0	1 sincerity
250	60	4.2	4.11	.05	.17	.8	0	.9	0	2 oral fluency
256	60	4.3	4.22	.22	.17	.7	−2	.7	−1	3 pronunciation
257	60	4.3	4.23	.24	.17	.9	0	.9	0	4 eye contact
236	60	3.9	3.88	−.36	.17	1.2	1	1.2	1	5 facial expression
243	60	4.1	4.00	−.15	.17	1.2	1	1.2	1	6 grammar
236	60	3.9	3.88	−.36	.17	.7	−1	.7	−1	7 originality
244	60	4.1	4.01	−.12	.17	1.0	0	.9	0	8 content
246	60	4.1	4.04	−.07	.17	.9	0	.9	0	9 written fluency
235	60	3.9	3.87	−.39	.17	1.1	0	1.2	0	10 evidence
180	60	3.0	3.02	.72	.19	1.3	1	1.4	1	11 holistic eval
239.9	60.0	4.0	3.95	.00	.17	1.0	−.1	1.0	−.2	Mean (Count : 11)
20.5	.0	.3	.32	.32	.01	.2	1.2	.2	1.2	S.D.

RMSE (Model) .17 Adj S. D. .27 Separation 1.56 Reliability .71
Fixed (all same) chi-square : 34.6 d. f. : 10 significance : .00
Random (normal) chi-square : 9.7 d. f. : 9 significance : .37

6, whereas student 4 is the poorest. The fit statistic shows that all the students fit the model except student 6, whose infit and outfit statistic scores are over the acceptable range (0.6-1.4). Whether the high ability of student 6 is related to the misfit result is not clear because student 8, whose ability is the highest, does not have any unexpected responses in the table. However, it is possible that student 6 could have brought about the unexpected score because of his relation to other factors, such as those in the third case of item 5 in Table 1.

The Separation index (2.66) of this measure is acceptable as an indicator for separating students because it exceeds the minimum requirements of an acceptable score of 2.0. It can be said that this presentation test was able to separate the students reasonably.

Next let us investigate the item measurement in Table 4.

Table 4 is the item measurement report. The fit statistic proves that all the items are functioning well within the acceptable range. It can be said that on the whole, all the items fit the model. The measure column suggests that the easiest item is number 11 (holistic evaluation) while the hardest are 5 (facial expression) and 7 (originality). One explanation for the results obtained in item 11 is that a 4-point scale is used only for this item (holistic evaluation), so raters' judgments do not have a wide distribution within this scale. An explanation for item 5 (facial expression) is that students can misinterpret facial expressions, even in Japanese conversations, because of the classroom environment. An explanation for item 7 (originality) is that students were assigned to use the target of the chapter. This limited their choices for freely expanding their ideas, even

Table 4' All Facet Vertical Rulers

Measr	+raters	+students	+items	S.1	S.2
+	2 +	+	+	+ (6)	+ (4) +
				5	
		8			
		6		–	3
+	1 + 1	+ 12	+	+	+ +
		5 9			
			holistic evaluation		
	4	1			
		7			
		3		4	
			eye contact pronunciation sincerity		–
		11			
*	0 *	* 2	* oral fluency	*	* *
			content written fluency		
		10	grammar		
		4			
	3 5		evidence facial expression originality		
				–	
	2				
+	–1 +	+	+	+ (2)	+ (1) +
Measr	+raters	+students	+items	S.1	S.2

though they were allowed to choose their own topics. Furthermore, students tend not to stand out among peers in class by doing extremely original things.

Let us look at the separation index of 1.56, which is below 2.0 (the suggested point, initially). It may be that some of the items do not function well in generating the expected distribution of students on the scale. This is probably because the number of items is not great enough to differentiate enough among the students' various abilities, so that some extremely good or extremely poor students were not measured well by these items.

Let us look at All Facet Vertical Rulers in Table 4'.

It is clear that the columns of students and items in Table 4' constitute data that confirm what was said above. The students are more widely dispersed on the scale than are the items. On the whole, however, all the items function well in measuring these 12

Table 5a Category Statistics.

Model = ?,?,1-10,R6

DATA Category Score	Counts Used	Cum. %	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	STEP CALIBRATIONS Measure	S. E.	EXPECTATION Measureat Category	−0.5	MOST PROBABLE from	THURSTONE THRESHOLD at	Cat PEAK Prob
2	11	2%	2%	−.10	−.62	1.2			(−4.13)		low	low	100%
3	147	25%	26%	−.21*	−.19	1.0	−3.01	.31	−1.79	−3.18	−3.01	−3.08	63%
4	270	45%	71%	.33	.38	.9	−.52	.10	.38	−.61	−.52	−.56	56%
5	116	19%	91%	.98	.99	1.0	1.53	.11	1.86	1.16	1.53	1.28	38%
6	56	9%	100%	1.69	1.51	.8	2.00	.16	(3.39)	2.70	2.00	2.36	100%
									(Mean)		(Modal)	(Median)	

Table 5b Category Statistics.

Model = ?,?,11,R4

DATA Category Score	Counts Used	Cum. %	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	STEP CALIBRATIONS Measure	S. E.	EXPECTATION Measureat Category	−0.5	MOST PROBABLE from	THURSTONE THRESHOLD at	Cat PEAK Prob
1	1	2%	2%	−.48	.25	.7			(−3.45)		low	low	100%
2	16	27%	28%	1.02	.65	1.8	−2.34	1.02	−.97	−2.46	−2.34	−2.39	66%
3	25	42%	70%	1.19	1.21	1.1	.47	.32	1.20	.22	.47	.33	50%
4	18	30%	100%	1.58	1.83	1.3	1.86	.32	(3.10)	2.27	1.86	2.04	100%
									(Mean)		(Modal)	(Median)	

students.

Let us take a look at the functioning of the rating items (1-10) in Table 5a.

Table 5a shows the category (rating items) statistics. Items 1-10 were rated on scales ranging from 1-6, although scale 1 was not used at all. The scores in the outfit column indicate that all the remaining scales (2-6) were reasonably used and that there were no misfitting scales among them. It should be noted that the lowest category (rating item) is never used by the raters. This is typical behavior, especially in a classroom situation. Peer students and teachers tend to avoid the lowest scale because they do not want to hurt others by giving them disappointing scores, even if the raters are anonymous. Therefore, we still need this unused bottom scale for the sake of students and teachers in a classroom setting.

Let us now look at the functioning of the rating item (11) in Table 5b.

Table 5b presents another category statistic for item 11, which was rated on a 1-4 point scale. Perhaps because only four different categories were used, category 2 exhibited a significant misfit as seen in the outfit statistic column. This probably caused the unexpected response in Table 1, where, as was pointed out, the value for item 11 was surprising, and the student whose expected score was 3.7 was placed in category 2. It is not clear what kind of complex relationship exists between category 2 and the three facets (rater, student and item), but category 2 has something to do with the unexpected

score. In this way, we can explore the integrated relationship between unexpected scores and category statistics.

5.2. How can the test be improved by taking into consideration the Rasch-based analyzed data?

Table 6 raters Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S. E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N raters
609	131	4.6	4.67	1.12	.11	1.3	2	1.3	2	1 1
447	130	3.4	3.43	−1.03	.13	1.4	2	1.3	2	2 2
498	132	3.8	3.75	−.36	.12	.6	−3	.6	−3	3 3
576	132	4.4	4.36	.67	.11	.9	0	.9	0	4 4
495	132	3.8	3.73	−.40	.12	.7	−2	.7	−2	5 5
525.0	131.4	4.0	3.99	.00	.12	1.0	−.4	1.0	−.4	Mean(Count: 5)
58.9	.8	.4	.46	.78	.01	.3	2.5	.3	2.4	S. D.

RMSE (Model) .12 Adj S. D. .77 Separation 6.49 Reliability .98
Fixed (all same) chi-square: 218.3 d. f.: 4 significance: .00
Random (normal) chi-square: 4.0 d. f.: 3 significance: .26

Table 7 students Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S. E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu students
221	55	4.0	3.98	.57	.18	1.4	1	1.4	2	1 1
205	55	3.7	3.69	.03	.19	.8	−1	.7	−1	2 2
217	55	3.9	3.91	.44	.18	.9	0	1.0	0	3 3
196	55	3.6	3.54	−.29	.19	.8	0	.8	0	4 4
221	54	4.1	4.04	.69	.18	1.0	0	1.0	0	5 5
237	54	4.4	4.36	1.19	.18	1.5	2	1.6	2	6 6
218	55	4.0	3.93	.47	.18	1.0	0	.9	0	7 7
246	55	4.5	4.47	1.35	.18	.7	−2	.7	−2	8 8
227	55	4.1	4.09	.76	.18	.8	−1	.8	−1	9 9
197	55	3.6	3.55	−.25	.19	1.2	0	1.1	0	10 10
206	55	3.7	3.71	.07	.19	.9	0	.8	−1	11 11
234	54	4.3	4.33	1.10	.18	.8	−1	.8	0	12 12
218.8	54.7	4.0	3.97	.51	.18	1.0	−.2	1.0	−.3	Mean(Count: 12)
15.1	.4	.3	.30	.52	.01	.3	1.3	.3	1.4	S. D.

RMSE (Model) .18 Adj S. D. .49 Separation 2.67 Reliability .88
Fixed (all same) chi-square: 96.3 d. f.: 11 significance: .00
Random (normal) chi-square: 11.0 d. f.: 10 significance: .36

In order to simplify the statistical interpretation of the test results, using the Rasch model, it is theoretically possible to delete the misfitting items, students, or raters. Then what is left could be regarded as the modified test item. However, deleting raters is not as easy as deleting students or items because the number of raters is usually not great. Therefore, even when only one rater is deleted, the effect on the whole can be disproportionately significant. Accordingly, the deletion of raters should be a last resort for

Table 8 items Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S. E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu items
256	60	4.3	4.21	.23	.17	.9	0	.9	0	1 sincerity
250	60	4.2	4.11	.06	.17	.9	0	.9	0	2 oral fluency
256	60	4.3	4.21	.23	.17	.7	−1	.7	−1	3 pronunciation
257	60	4.3	4.23	.26	.17	1.0	0	.9	0	4 eye contact
230	59	3.9	3.84	−.43	.18	1.1	0	1.1	0	5 facial expression
243	60	4.1	3.99	−.15	.17	1.2	1	1.2	1	6 grammar
236	60	3.9	3.88	−.36	.18	.8	−1	.8	−1	7 originality
244	60	4.1	4.01	−.12	.17	1.0	0	.9	0	8 content
240	59	4.1	4.00	−.13	.17	.8	−1	.7	−1	9 written fluency
235	60	3.9	3.87	−.39	.18	1.2	0	1.2	1	10 evidence
178	59	3.0	3.05	.78	.19	1.3	1	1.3	1	11 holistic evaluation
238.6	59.7	4.0	3.95	.00	.18	1.0	−.1	1.0	−.2	Mean(Count: 11)
21.1	.4	.3	.31	.34	.01	.2	1.1	.2	1.1	S. D.

RMSE (Model) .18 Adj S. D. .30 Separation 1.69 Reliability .74
Fixed (all same) chi-square: 38.7 d. f.: 10 significance: .00
Random (normal) chi-square: 9.8 d. f.: 9 significance: .37

improving statistical results.

Then, what can be done to improve the situation? First, let us look back at Table 1 and examine the details. Three responses (in which a rater, an item and a student are interrelated) are determined to be unexpected. Now, let us delete various combinations of three facets in three cases: case one (rater 1, student 12, item 11), case two (rater 2, student 5, item 9), and case three (rater 2, student 6, item 5).

Tables 6, 7, and 8 show the results. Table 6 indicates no misfitting rater in the column of infit and outfit statistic. Table 7 shows one misfitting student (student 6) in the column of infit and outfit statistic. Table 8 presents no misfitting item in the column of infit and outfit statistic.

As far as items are concerned, all of them function well with 5 raters to measure the students. We did not find any misfitting items in this small sample. Since one of the purposes of this research is to calibrate items for an item bank, these items in theory can be kept in the item bank as mentioned later.

On the other hand, as Student 6 turned out to be misfitting, it is worthwhile to examine why this student behaved idiosyncratically. Although Student 6 should be deleted for statistical analysis, he may provide some information about the test takers' internal behavior.

5.3. How can item banks be constructed?

Item banks are collections of test questions that are stored in special computer

programs where storage is structured or organized according to the codes assigned by users. (Rudner 1998).

One of the code sets includes item characteristics (item difficulty in this present case). The determination of item characteristics can be done in one of two ways: either by Classical Test Theory or by Item Response Theory(IRT).

Once all the items are calibrated and the difficulty of each item is determined each item can be put on the continuum of the scale according to their logit scores (difficulty levels). These items along with a task (speech presentation) can be stored as items in a bank.

The data in the present research has already been calibrated by using the Rasch statistical model which is one of the item response theories. Since the items are calibrated, we can store those items in the bank. This stage is called the deposit stage, where items are entered into special computer files. This stage is followed by the bank stage where items are stored in suitably labeled computer files. The bank stage is in turn followed by the withdrawal stage, where items are selected from the bank based on specific needs to measure test takers' ability more accurately. (cf. Rudner 1998)

In the case of the present research, since all the items have already been calibrated by the IRT based Rasch model, it can be said in theory that we can store these items in the bank stage through the deposit stage and wait for an occasion where they will be selected to match the test taker' needs or to measure their ability. However, in practice, there are some necessary procedures to make the item bank more reliable, such as increasing the number of test takers (at least 100 students), or training the raters.

Once the initial bank has been established, an advantage of calibrated item banks is in the ease of test development. Teachers withdraw from the bank those items most suitable, in terms of difficulty level, to measure the students' ability. On the basis of the test results, teachers gain greater insight into the learning process of their students. Eventually, this will be reflected in the curriculum. (Rudner 1998).

6. Implications and conclusions

Some conclusions can be drawn. Firstly, the Rasch based analysis provides us with 1) the relationship among three facets of peer evaluation (raters, students, items), 2) the rater severity and fit statistic, 3) students' ability and fit statistic 4) item difficulty and fit statistic and 5) an assessment of the functioning of rating categories. With all or some of these pieces of information, the three facets of a test can be thoroughly investigated individually--a task that would not be possible in a traditional test analysis.

Secondly, a test can be improved by examining the fit statistic (misfit items) statistically. Also, the test can be improved by having a discussion with raters or students when they are misfits.

Lastly, the Rasch model can greatly help us construct item banks in the process of item calibration. Furthermore, a carefully calibrated item bank can make a great contribution to the future use of presentation assessment protocols as well as to curriculum development.

Note : This research was supported in part by Tokyo Keizai University under Research Grant CPU04-00.

7. Bibliography

Linacre, John M. (1989, 1994). *Many-Facet Rasch Measurement*. Chicago : MESA Press.

Oscarson, Mats (1997). *Self-assessment of foreign and second language proficiency*. In C. Clapham and D. Corson (Eds). Encyclopedia of Language and Education, Volume 7 : Language testing and assessment, 175-187.

Rudner, L. (1998). *Item banking*. ERIC/ AE Digest Series EDO-TM98-05.

Tatum, Donna (1997). *Meaningful Measurement : Competency Map--An Item Bank for Speech Evaluation*. Chicago, Illinois. Meaningful Measurement.